Testing the Hourglass Model of Vertebrate Development: Methods for Developmental Transcriptomic Meta-Analyses

Pranav Subramanyam Bhamidipati

BCH 379H
Special Honors in the Department of Biochemistry
Submitted to fulfill the Plan II Honors Program thesis requirement
The University of Texas at Austin

May 9th, 2017

Hans Hofmann
Department of Integrative Biology
Supervising Professor

Jeffrey Barrick
Department of Molecular Biosciences
Biochemistry Honors Advisor

ABSTRACT

Author: Pranav S. Bhamidipati

Title: Testing the Hourglass Model of Vertebrate Development: Methods for Developmental Transcriptomic Meta-Analyses

Supervising Professors: Drs. Hans Hofmann and Jeffrey Barrick

Evolutionary variation is responsible for a broad diversity among organisms and driven mechanistically by conserved developmental processes. The hourglass model of development posits that all organisms sharing a lineage (phylum) undergo a period of similarity in during an intermediate period of embryonic development, with increased divergence at the beginning and end. Quantitative assessment of this hypothesis is now feasible with recent technological advances in gene expression profiling and the widespread availability of gene expression profiling data in public repositories. However, no standards for best-practices have been established to guide meta-analysis of long time-series transcriptomic data across taxa. We are conducting a meta-analysis of gene expression profile studies in five vertebrate species (Danio rerio, Gallus gallus, Mus musculus, Xenopus laevis, and Xenopus tropicalis), from the beginning to the end of development, to test the existence of an hourglass pattern and probe its molecular mechanisms. In order to achieve this, we identified and addressed three principal challenges – batch effects, multiple profiling platforms, and broad sampling/low sample size. First, the collected data were manually curated and tested for sources of variation to minimize batch effects caused by differing methodologies. Next, a pipeline of data transformations was devised to integrate data from microarray (MA) and RNA-seq (RS) profiling techniques in X. tropicalis. Then, four naïve descriptive metrics (CV, mean FC, max FC, and max FC/%T) were evaluated as selectors for genes or orthologous gene groups (OGGs) showing important temporal expression patterns. These metrics were then evaluated for use in selecting important developmental genes/OGGs to reduce the ratio of factors to samples. PCA results indicated that curation successfully resulted in a meta-analysis with no detectable batch effects. The MA-RS integration pipeline, on the other hand, showed poor effectiveness in eliminating batch effects from profiling platform and limited translation to other genera. Expression importance metrics appear roughly equal in their discrimination of important patterns. Preliminary data show evidence of an hourglass pattern of gene expression, and importance metrics are being applied in tandem to study the role of developmentally important genes in generating this pattern.

Introduction

Organisms display an astonishing variety of phenotypes, yet the developmental underpinnings of diversity are relatively conserved at the genetic level, suggesting a common mechanistic

program underlying the evolution of phenotypic diversity. Since the 1820s, it has been observed that developing animals undergo a "phylotypic period," or a period in development during which embryos in a given lineage exhibit greater morphological similarity than at the beginning or end of development (Fig. 1A). [1–5] This "hourglass" model of development is supported by recent comparative transcriptomic analyses [6,7] and could result from greater genetic interaction (e.g. pleiotropy and/or epistasis) in the phylotypic period.[3,4] However, an hourglass pattern of variation in gene expression has not been robustly demonstrated, and alternative hypotheses such as early conservation should also be considered (Fig. 1B).

To test the hypothesis of an hourglass pattern, we are performing a meta-analysis of publicly available transcriptomic data collected from five vertebrate species (*Danio rerio, Gallus gallus, Mus musculus, Xenopus laevis, and Xenopus tropicalis*) from early to late embryogenesis. Meta-analyses across both developmental and evolutionary time are uncommon, and few resources are available to carry them out. We addressed three significant methodological challenges in order to fully utilize the available data to address questions in organismal evolution and development. We endeavored to (1) curate the use of data sets to minimize batch effects between studies, (2) resolve differences in the technique/platform used for transcriptome profiling between studies, and (3) maximize the ratio of sampled factors (e.g. genes) to sample size (e.g. biological replicates).

Background of Methodology

Meta-analyses of gene expression data face many challenges, the most apparent of which is susceptibility to batch effects. RNA sample quality, processing, and storage can vary

significantly between gene expression studies and introduce systematic variations that hinder sensitivity in gene expression measure ements.[8,9] On the scale of meta-analysis, methodological differences can introduce significant batch effects between data sets that harm accuracy, sensitivity, and reproducibility.

Another significant obstacle to transcriptomic meta-analysis is the use of two disparate gene expression profiling technologies, oligonucleotide microarrays and next-generation sequencing (RNA-seq). Microarrays measure fluorescence signal from antiparallel complementary binding of transcripts and oligonucleotide probes. Arrays are used extensively but have various limitations. Optical noise and saturation restrict their dynamic range, cross-hybridization poses challenges to probe set design, gene sampling is heavily dependent on genome annotation, and resolution is generally insufficient to differentiate gene isoforms.[10] The development of RNA-seq has enabled direct sequencing and quantification of cDNA to mitigate many of these problems. However, no computational tools have been built to compare and integrate microarray and RNA-seq data. It is unclear to what extent these two methods are comparable, but integration may be possible under certain restraints. Concordance in signal-response (that is, similarity in measured expression signal in response to the same change in expression) between microarray (MA) and RNA-seq (RS) is higher for expression values above the median level of transcript abundance.[11,12] A method to integrate data from these technologies could take advantage of these trends to maximize coherence across data sets and between platforms.

A third problem that plagues many gene expression profile studies, and systems biology studies in general, is the "curse of dimensionality," or the "$p >> n$ problem." Gene expression

analyses tend to have low sample sizes because of the high cost per sample. Often, these studies observe many more factors (p) than samples (n).[13] Correlational analyses may become unreliable from lack of statistical certainty, and model-based methods, such as inference of gene regulatory networks, particularly suffer. Apart from reducing the cost-per-sample of profiling, solutions to this problem could involve creative methods of sample pooling and more selective profiling of the transcriptome.

Methods

Curation of Collected Transcriptomics Data

To mitigate batch effects from multiple data sources, relatively strict criteria were applied to select only data sets that were well-characterized and amenable to meta-analysis, and indistinguishable from other conspecific data sets using naïve clustering. Data sets were curated using a number of factors. First, publications that provided sparse or unclear descriptions of metadata and data collection were removed. After pre-processing, data sets were excluded if they showed a low coverage of genes in the genome relative to other conspecific data sets from the same platform. Data sets that covered a small number of developmental time points (~1-2) were also excluded to avoid introducing batch effects indistinguishable from biological variation between time points. Histograms of gene expression at each time point and across all time points were compared qualitatively, and data sets with grossly different signal response/dynamic range of expression were excluded. Finally, PCA analyses were run using

expression data for all species at all time points. Data sets that clustered separately from other data sets in the same species (i.e. introduced a batch effect) were excluded.

Description of Data Sets

Eight gene expression profile data sets were collected from publicly available sources and repositories. In *Danio rerio,* microarray transcriptomic data for 2 replicates of 10 developmental time points (shield, 75% epiboly, 90% epiboly, bud, 5-somite, 14-somite, prim-5, 32 hours post-fertilization (hpf), long-pec, and 4 days post-fertilization (dpf); alternatively, 6 hpf, 8 hpf, 9 hpf, 10 hpf, 12 hpf, 16 hpf, 24 hpf, 32 hpf, 2 dpf, and 4 dpf) were downloaded from the EMBL-EBI ArrayExpress repository (Accession: E-TABM-33). RNA-seq data for 1 pooled replicate of 9 time points (2 hpf, 4 hpf, 5 hpf, 16 hpf, 36 hpf, 48 hpf, 60 hpf, and 72 hpf) were downloaded directly from the publisher's website.[14,15] *D. rerio* expression data at the zygote developmental stage (0.25 hpf) was excluded because of likely abundance of maternal transcript, and time points after 4 dpf were excluded due to completion of the developmental program.[15,16] In *Gallus gallus*, microarray data for 2 replicates of 15 time points (Hamburger Hamilton stages 1, 2, 4, 6, 8, 9, 11, 14, 16, 19, 24, 27, 32, 34, and 38) were downloaded from ArrayExpress (Accession: E-MTAB-366).[6] *Mus musculus* microarray data for 2 replicates of 8 time points (Theiler stages 11, 13, 15, 17, 22, 24, and 26) were downloaded from ArrayExpress (Accession: E-MTAB-368), and data for 3 replicates of 11 time points (Theiler stages 1, 9, 11, 13, 16, 19, 21, 22, 23, 25, and 27) were downloaded from the NCBI Gene Expression Omnibus, or GEO (Accession: GSE39897).[6,17] In *Xenopus laevis* and *Xenopus tropicalis*, microarray data for 15 time points (stages 2, 8, 9, 10, 12, 13, 14, 16, 18, 20, 23, 25, 30, 33, and 40) were downloaded from GEO (Accession: GSE27227).[18] In

Xenopus tropicalis, RNA-seq data for 22 time points (2, 8, 9, 10, 11, 12, 11-12, 13-14, 15, 16, 16-18, 19, 20-21, 22-23, 24-26, 28, 31-32, 33-34, 38-39, 40, 41-42, and 44-45) were downloaded from GEO (Accession: GSE37452).[19] The sample size of this data set varied from 1-3 replicates (See Supplemental Table 1). See Supplemental Table 2 for the elapsed real time at each time point for all data sets.

Data extraction and pre-processing

For all data sets, expression data were originally collected and published in one of three formats: .CEL file (Affymetrix microarrays), Agilent results file (Agilent microarrays), or data table (both RNA-seq studies). All data transformations and analyses were performed with the statistical computing software and programming platform R (https://www.r-project.org). Affymetrix and Agilent data were imported using the R packages *simpleaffy* and *limma*, respectively.[20,21] RNA-seq data sets were imported from data tables with genes identified by ENSEMBL gene ID and gene expression given in reads per kilobase per million mapped reads (RPKM). For both microarray platforms, pre-processing consisted of RMA background correction with quantile normalization.[22] For Affymetrix data sets, the package *biomaRt* was used to assign each probe set to its corresponding ENSEMBL gene ID(s). (This information was automatically attained by *limma* for Agilent data sets.) For Affymetrix and Agilent data, probe sets that mapped to multiple genes or no genes at all were excluded from further analysis. The signals of all probe sets mapping to the same gene were averaged to give each gene a singular expression value. All expression values were Log2-transformed, or transformed using the

function $log_2(x)$. To find average gene expression during a given time point of a species' development, mean expression was computed across all samples of that time point.

Ortholog Calling and Orthologous Gene Group Assignment

The program OrthoMCL was used to classify all genes across all 5 species into orthologous gene groups (OGGs), each containing paralogs (conspecific homologous genes) and orthologs (interspecific homologous genes). This homology is calculated based on BLASTP, a protein sequence homology algorithm. (These steps were performed by Dr. Rebecca Young-Brim.) The resulting clusters assigned each gene to one OGG, and each organism's OGG "expression" was computed as the average of all paralogs.

Alignment of Microarray and RNA-seq Distributions

To align the frequency distributions of the paired RNA-seq and microarray data sets in *X. tropicalis* at the third quartile (Q3), a sequence of transformations were applied. First, all RNA-seq expression values below $Log_2(RPKM) = -4$ were assigned a value of -4. Then, the following correction factor was applied,

$$F = Q3(MA) - Q3(RS)$$

$$RS^* = RS + F$$

where F is a scalar correction factor, Q3(x) computes the third quartile of a data table x, MA is a data table of microarray expression, RS is a data table of RNA-seq expression, and RS* is a data table of corrected RNA-seq expression. The resulting expression value will be henceforth referred to as "Log(expr)." Finally, a high-pass filter of Log(expr) > 10 was enforced to eliminate

below-threshold expression (threshold chosen arbitrarily). This method was then evaluated by least-squares regression of $RS{\sim}MA$ and calculation of the coefficient of determination (R^2).

Selection of Gene Importance

Gene (or OGG) importance was assessed using both descriptive and inferential methodologies.

Four descriptive metrics were measured: (i) coefficient of variation (CV), (ii) mean fold-change between time-points (meanFC), (iii) maximum fold-change between time-points (maxFC), and (iv) maximum fold-change over percent of developmental time (maxFC/%T). They are defined below.

Let there be a set of Log2-transformed expression values $E = \{e_1, e_2, \dots e_n\}$ that spans an ordinal sequence of time-points t: $\{t_1, t_2, \dots t_n\}$ denoting the stages of development.

(i) CV

$$\mathrm{CV(E)} = \frac{\mathrm{s(E)}}{\bar{\mathrm{x}}(E)} \times 100\%,$$

where

$$s(E) = standard\ deviation\ of\ E, \qquad \bar{x}(E) = mean\ of\ E$$

Let there be a function $FC(S)$, where $S = \{s_1, s_2, \dots s_n\}, n \geq 2$.

(ii) meanFC

$$FC(S) = \{fc_1 = |s_2 - s_1|, fc_2 = |s_3 - s_2|, \dots fc_{n-1} = |s_n - s_{n-1}|\}$$

(Fold-change is calculated by subtraction because the values are log-transformed.)

$$meanFC(S) = \frac{1}{n-1}\sum_{i=1}^{n-1} fc_i$$

(iii) maxFC

$$\text{maxFC(S)} = \max[FC(S)]$$

Let there be a numerical sequence of time-points $T = \{T_1, T_2, \ldots T_3\}$, measured in hours or days post-fertilization (hpf or dpf). Let there be a function

$$\%T = \frac{T}{\max[T]} \times 100\%$$

that measures the percent of developmental time elapsed at each time-point.

(iv) maxFC/%T

$$FC/\%T(S) = \left\{fct_1 = \left|\frac{s_2 - s_1}{\%T_2 - \%T_1}\right|, fct_2 = \left|\frac{s_3 - s_2}{\%T_3 - \%T_2}\right|, \ldots fct_{n-1} = \left|\frac{s_n - s_{n-1}}{\%T_n - \%T_{n-1}}\right|\right\}$$

$$\max FC/\%T(S) = \max[FC/\%T(S)]$$

Clustering of developmental time-points into periods

Adjacent time-points were clustered into groups of time-points, or "periods," using *k*-means clustering in R. Sums of squares error (SSE, or variance within cluster) was computed for k = 1-9, and no "elbow" effect (sharp drop in SSE at optimal k) was observed. Consequently, k = 3-6 were chosen for further analysis based on a minimum of 3 periods to observe an hourglass pattern and reduced numbers of time-points within each period for $k > 6$. All OGG expression data within a time period was pooled for comparison across species and between periods.

Results

Selected data sets describe transcriptomic variation due to evolution and development

Curation of a set of 20 publications streamlined the meta-analysis to 8 coherent and appropriate data sets that did not show evidence of batch effects, yet still sampled the full range of stages of development in each species. Only studies that spanned numerous time-points of development, retained a sufficient number of sampled genes after pre-processing, and originated from a publication that explicitly provided necessary metadata (e.g. sample size) were used in further analyses.

Quality and appropriateness of these data sets were assessed using principal components analysis (PCA) of all orthologous gene groups (OGGs) in all species over all time points. A data set was considered to introduce batch effects if it visually separated from other conspecific data sets in the first few principal components (PCs) of PCA. One outlying data set was found. The first principal component (PC1) identified *M. musculus* transcriptomic data for Theiler stages 2, 3, and 4 as significantly different from all other *M. musculus* time points (Figure 2A).[23] Indeed, this difference appeared greater even than that between *M. musculus* and some other species. When *M. musculus* data for TS2, TS3, and TS4 were removed from the PCA, no other data sets appeared to contribute significant batch effects (Figure 2B). PC1 described 83.02% of variation and primarily distinguished *Xenopus* spp. from other vertebrates. PC2 described 4.75% of variation and distinguished the rest of the species from one another. PC3 described 3.06% of the variation and spread out time points in chronological order, with few exceptions. Thus,

evolutionary divergence and progression through development appear to be the principal factors modulating temporal variation in gene expression in our meta-analysis.

Integrating Microarray and RNA-seq Data Sets

Approximately 1/3 of collected data originated from studies that used RNA-seq (RS) transcriptional profiling, while the rest were produced using microarray chips (MA). In order to boost the sample size of transcriptomic data that could be pooled for analysis, a computational pipeline was developed to attempt to integrate MA and RS transcriptional profiles. Based on the assumption of higher concordance above median expression and the bimodal shape of the *Xenopus tropicalis* gene expression frequency distribution, a pipeline of data transformations was applied to a pair of *X. tropicalis* MA and RS data sets at 12 shared time-points of development (stages 2, 8, 9, 10, 11-12, 13-14, 16-18, 20-21, 22-23, 24-26, 33-34, and 40) to isolate genes in this region of higher concordance.

Within each time point, the transcriptomes of *X. laevis* and *X. tropicalis* were found to have a bimodal frequency distribution, the mean and median falling between the two modes (Figure 3A). A series of data transformations were performed to align the RS frequency distribution to the same scale as MA expression values (henceforth denoted as Log(expression), or Log(expr)) and select the higher mode for analysis. This pipeline aligned the higher mode of the MA and RS distributions (Figure 3B). A high-pass filter was then applied to remove the unaligned lower modes. This improved correlation between MA and RS in *X. tropicalis* Stage 10 ($R^2 = 0.949$ to $R^2 = 0.980$) and shifted the slope of the least-squares regression line from 0.929 to 0.994, very close to

the ideal slope of 1 (Figure 3C). These results suggest a possible future for RS-MA data integration using the pipeline (summarized in Figure 3D).

However, a PCA performed post-hoc on the aligned MA and RS data showed that post-correction RS and MA expression contributed a great amount of variation (Figure 3E). In PC1, PC2, and PC3, expression data at stages analyzed by different platforms showed dissimilar clustering, to a similar extent as differences between species. Additionally, the bimodal expression pattern required for application of a high-pass filter was not observed for any other species in the meta-analysis (data not shown). Thus, even after adjusting for disparities in dynamic range, RS and MA data vary significantly from one another, and the described pipeline may have minimal usefulness even if this issue were to be fixed.

Scoring gene/OGG importance with naïve descriptive metrics

Preliminary results suggested that many genes in the genome do not significantly modulate their expression over the course of embryonic development (data not shown). Thus, isolation of the genes most important to the developmental program may both reduce the p/n ratio and increase the signal-to-noise ratio of observed changes in the overall transcriptome. Given a large number of genes relative to sample sizes within species (Table 2) and many orthologous gene groups (p = 2529 OGGs), a necessity was foreseen for reducing the number of genes in the analysis while maintaining or enriching the proportion of developmentally significant genes.

Towards such an effort, four naïve descriptive metrics were conceived to compute the importance of genes in development solely from their temporal expression pattern – CV, Mean FC, Max FC, and Max FC/%T.These metrics were developed and evaluated based on the model

that an important change in gene expression manifests as an impulse response or sustained response in expression.[24] The sustained response pattern has particular relevance to development, as changes in gene expression states underlie changes in cell state over the course of development. See Fig. 4A for examples of both temporal expression patterns using the *X. tropicalis* genes ENSXETG00000012655 and ENSXETG00000024597.

These four metrics assay different expression patterns. To test the extent of consensus, all four metrics were computed for *X. tropicalis* OGG expression. Histograms of each metric are strikingly similar, showing pronounced right skew of the mean (Fig. 4B-E). There is, however, no clear separation of sub-populations of OGGs undergoing different expression patterns. To assess the extent to which the four metrics agree in their scoring of OGGs, the overlap of OGGs scoring above a threshold quantile value h was calculated between each pair of metrics for h = 0.50, 0.75, and 0.90 (Figs. 4F-H). For all quantile thresholds, there did not appear to be a pair of metrics with significantly less or more overlap. These results suggest that all four metrics capture a similar proportion of time-variant genes, and no one metric is noticeably superior or similar to any others.

Preliminary analysis suggests an hourglass pattern in interspecific variation

The large number of OGGs relative to time-points is being addressed by clustering time points into periods for cross-species comparisons. Clustering into 3 periods resulted in a statistically significant pattern of descending gene expression correlation ($p < 0.05$, data not shown), supporting the alternative hypothesis of early conservation of variation. On the other hand, clustering into 4-6 periods consistently presented an hourglass pattern but did not cross the

significance threshold $\alpha = 0.05$. Other preliminary data (not shown) have suggested that the hourglass pattern is indeed present but obscured by a large amount of coincidental correlation between genes, a by-product of the large number of sampled genes and a preponderance of genes not undergoing significant changes over time.

Discussion

Vertebrates exhibit considerable diversity in adult phenotypes and reproductive strategies; nonetheless, they appear to share anatomic and transcriptomic similarity at an intermediate "phylotypic" period of development (Fig. 1A). This qualitative observation of a developmental "hourglass" has been a standing question in embryology since its 19th century conception, but only recently has it become quantitatively testable against alternative hypotheses (Fig. 1B). The emergence of gene expression (transcriptome) profiling and cross-species gene orthology technologies has resulted in a wealth of publicly available and comparable gene expression profiling studies in different vertebrates. However, few resources exist to facilitate meta-analysis of transcriptomic data across species or over long time-scales.

To validate the hourglass pattern and investigate its underlying mechanisms, 8 publicly available data sets were downloaded from online repositories representing the gene expression profiles of *Danio rerio, Gallus gallus, Mus musculus, Xenopus laevis,* and *Xenopus tropicalis* from early to late embryonic development. In order to test the hypothesis of an hourglass pattern, three main methodological challenges of transcriptomic meta-analyses were undertaken – curation of data sets to reduce batch effects, integration of data produced by disparate profiling platforms, and maximization of the sample size relative to the number of sampled factors.

Curation is an essential first step of transcriptomic meta-analysis

Publicly available data sets from twenty publications were downloaded and were first subject to stringent quality control curation. Data sets were curated with respect to study design, methods documentation, coverage of the transcriptome, number of sampled time points, and coherence with other collected data. After pre-processing and orthology, coherence was determined by performing principal components analysis (PCA) of all developmental stages in all species to determine whether batch effects contributed a significant amount of variation relative to species and time point of development. *M. musculus* data from one publication in particular, Maekawa, et al. (2007),[23] clustered separately from other data sets of the same species and contributed a batch effect with effect size comparable to interspecies variation (Fig. 2A). This data set described otherwise unsampled stages of *M. musculus* development (Theiler stages 2, 3, and 4), but was removed nevertheless. PCA of the 8 data sets that satisfied all the applied criteria indicated that the greatest sources of variation in the data were species (encompassing PC1 and PC2) and stage in development (PC3) (Fig. 2B). Thus, these quality control steps led to strong, clear signals for the variables most relevant to addressing the hypothesis. These results indicate that data curation is an essential starting point for transcriptomic meta-analysis, particularly involving multiple axes of variation (e.g. evolutionary divergence and developmental time).

Concordance between microarray and RNA-seq platforms does not yield compatibility

The two predominant gene expression profiling technologies, microarray (MA) and RNA-seq (RS), have been extensively studied, rarely compared, and never integrated to enable simultaneous co-analysis. Ideally, expression values from the two platforms should show a high linear correlation with a slope of ~1 to suggest good concordance in expression and differential expression. A computational pipeline was implemented to integrate paired RNA-seq and microarray data sets based on the assumptions of a bimodal frequency distribution of gene expression (observed in *Xenopus* spp.; Fig. 3A) and higher concordance at above-median transcript abundance levels. After enforcing a lower limit of RS expression and aligning the two distributions at the third quartile (Q3, the central tendency of the above-median mode) by adding a scalar correction factor (Fig. 3B), a filter was applied to isolate the higher mode of the bimodal distribution. The approach showed some success in improving correlation and the slope of linear regression (Fig. 3C; pipeline summarized in Fig. 3D). However, when added to the meta-analysis after gene orthology computation, the pipeline-transformed data failed to remove the variance between MA and RS. When added to the MA data, the RS data introduced significant batch effects visible in the first principal component of interspecies PCA analysis, indicating that it contributes significantly to variation and coheres poorly with the other data sets (Fig. 3E). Furthermore, the general approach was dependent on the bimodal frequency distribution of the developing *Xenopus* transcriptome and therefore would be limited to similarly bimodal expression distributions. Bimodal distribution of expression was not observed for any other species in the meta-analysis. Due to the much broader sample space and dynamic range of sequencing-based profiling methods over oligonucleotide probes, cross-platform integration can also compromise the number of sampled genes observed by RS

profiling and curtail measurements at lower levels of transcript abundance. Overall, even if this pipeline could be modified to successfully perform MA-RS integration, its disadvantages may outweigh the advantages.

Gene selection could boost power and reduce noise for time-series study of systems data

Studies of large systems across the sciences are plagued by numerous interacting factors sampled by too few observations to draw reliable conclusions. This "curse of dimensionality" manifests in gene expression profiling as a small number of observations *n* of a large number of factors *p* (genes), a result of the costliness of profiling. Small sample sizes hinder the reliability of expression correlation and gene co-expression/network analyses alike. This challenge is complicated by the fact that development is a sequence of events that spans a considerable time, and resources for classifying long time-series expression patterns are sparse and computationally intensive.[25] To address these problems in a time-series context, four descriptive metrics (CV, Mean FC, Max FC, and Max FC/%T) were devised to score genes for importance, based on an assumption that important genes undergo an impulse or sustained response in temporal gene expression (See Fig. 4A for examples).[24] Each of the descriptive methods captures different characteristics of temporal expression.

CV, or coefficient of variation, computes a ratio of variance to mean and does not utilize temporal information. It thus should sensitively identify gross variation but may miss important temporal patterns such as a transient impulse or a change in expression near the boundaries (beginning or end) of a time-course. Mean FC computes the mean fold-change in

expression (FC) between adjacent time-points. Unlike CV, it should bring temporal patterns into register but may similarly miss transient impulses or boundary changes in expression. It may also be sensitive to noise in gene expression, which can vary across the genome due to differing promoter architectures and stochastic factors in activation.[26] Max FC computes the maximum FC across the time-course. While it is very robust to noise and captures significant time-dependent expression spikes/drops, it may be fooled by the heterogeneous rate of sampling over developmental time.

Because events of interest in development may be sampled heterogeneously rather than continuously, large periods of time may exist between adjacent sampled time-points. As a result, a minimal rate of change in expression may result in a high fold-change in expression when sampled at two distant time-points. To avoid this, FC was divided by the time elapsed (in percent of total time of development, or %T), giving a value identical to the slope of the curve Log_2(expression) vs. % development time. The metric max FC/%T computes the maximum of this value across development. Overall, max FC/%T accounts for sampling rate while retaining sensitivity to large changes in expression. However, it is sensitive to noise for very closely-spaced time-points, particularly common near the beginning of development.

Each metric produced a distribution with a pronounced right skew when applied to OGG data from *X. tropicalis* (Fig. 4B-E), suggesting that they may be useful in identifying OGGs with more time-variant and developmentally important expression patterns. However, no metric showed an ability to resolve multiple sub-populations of OGGs based on expression pattern, suggesting that either the expression patterns traced by OGG expression do not tightly follow a pattern

similar to the impulse/sustained response models or naïve metrics in general are insufficient to distinguish higher-order temporal expression patterns. There was significant pairwise overlap between metrics in the number of OGGs scoring above a threshold quantile *h* (Figs. F-H). Surprisingly, no pair of metrics showed significantly more or less pairwise overlap than the others for any of the thresholds tested (h = 0.50, 0.75, and 0.90), indicating that each metric is likely tracking a different aspect of "important" expression to an equal extent. While this result does not encourage special faith in any one metric, it does support their further use in tandem to isolate genes or OGGs that show significant expression patterns, in whatever form that might take.

Developed methods differ from previous cross-species transcriptomic studies

Previously, Irie and Kuratani (2011)[6] conducted a meta-analysis study evaluating the hourglass model. This study addresses many similar problems, and the differences between our approaches highlight particular challenges of gene expression studies across developmental and evolutionary time. In particular, a lack of a standard for best-practices makes curation and documentation practices critical to repeatability and translation. We eliminated *M. musculus* data from Maekawa, et al. (2007) (TS2, TS3, and TS4) due to batch effects (Fig.2A), while Irie and Kuratani (2011) did not report batch effects with this data set. Additionally, Irie and Kuratani made a transcriptomic data set for *Xenopus laevis* development available for download from a public repository, but the provided data format could not be converted to ENSEMBL gene IDs. Our analysis included time-points of *D. rerio* development up to day 4, the end-point of embryonic development,[16] while Irie and Kuratani curtailed their developmental series at day 3.

No significant batch effects remained after our additional curation steps, indicating that they may be a suitable basis for future best-practices.

When comparing time-points of development between taxa, Irie and Kuratani (2011) correlated expression at particular developmental stages across species (cleavage, blastula/shield, pharyngula, and latest) and demonstrated an hourglass pattern. While this approach may encourage comparison during similar points in development, it removes the majority of the available data from consideration during analysis and exacerbates the $p >> n$ problem. In our current preliminary analyses, we are pooling time-points from each species into 3-6 groups using k-means clustering, which reduces the p/n ratio by increasing the number of samples per period.

Notably, Irie and Kuratani's (2011) use of anatomical staging criteria also calls into question the merits of using gross, qualitative criteria to perform molecular, quantitative cross-species correlations. The "latest" developmental stage in their study represented the last stage sampled before the developmental program was completed, which may be highly variable across species. Thus, lower correlation observed at the "latest" stage may reflect poor comparability of time-points rather than evolutionary divergence. Furthermore, divergent species that evince some of the same anatomical features of developmental progress may not be executing the same developmental functions on a subcellular level. Identifying stages as identical that share a limited number of anatomical characters may inadvertently result in comparisons between species at stages that, in fact, are dissimilar. On the contrary, clustering time points into periods

does not necessarily presume that individual time-points are equivalent across species, but rather quantitatively infers periods of development and assumes a similar state or function.

Future directions

Preliminary results from interspecies correlation analyses suggest that an hourglass pattern most likely exists. However, the reliability and granularity of the pattern is subject to a tradeoff between the sample size within each cluster of stages and the total number of clusters. Currently, this problem is being addressed by selecting genes/OGGs significant to development using a combination of the aforementioned importance metrics. Going forward, these results will be repeated using the RNA-seq data sets. Methods should also be explored to enable the construction of gene regulatory networks to interrogate the topology of gene-gene interactions. These efforts will enable investigation of the mechanistic basis of the hourglass pattern and the vertebrate developmental program.

Acknowledgments

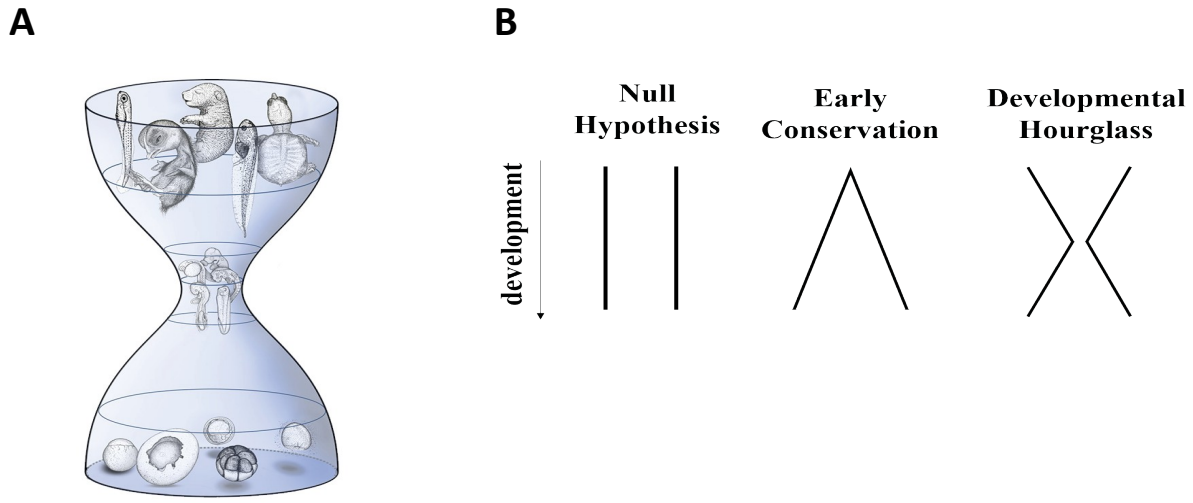
I would like to acknowledge Dr. Hans Hofmann for his guidance and wisdom and Dr. Rebecca Young-Brim for her mentorship. Data analyses in Figures 2B and 3A, gene orthology analysis, and k-means clustering were performed by Dr. Young-Brim. Data analyses in Figure 3E were performed by Paul Tee. I acknowledge Dr. Dennis Wylie and the UT Austin Center for Computational Biology and Bioinformatics for their bioinformatics consultation services. Funding for this work was provided by the BEACON Center for the Study of Evolution in Action and an Undergraduate Research Fellowship from the UT Austin College of Natural Sciences Office of the Vice President for Research.

References

1. von Baer, K. E. *Uber Entwickelungsgeschichte der Thiere*. (1828).
2. Medawar, P. B. The Significance of Inductive Relationships in the Development of Vertebrates. *Development* **2,** 172–174 (1954).
3. Duboule, D. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev.* **1994,** 135–142 (1994).
4. Raff, R. A. *The shape of life: Genes, development, and the evolution of animal form*. (University of Chicago Press, 1996).
5. Kalinka, A. T. & Tomancak, P. The evolution of early animal embryos: conservation or divergence? *Trends Ecol. Evol.* **27,** 385–393 (2012).
6. Irie, N. & Kuratani, S. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat. Commun.* **2,** 248 (2011).
7. Tena, J. J. *et al.* Comparative epigenomics in distantly related teleost species identifies conserved cis-regulatory nodes active during the vertebrate phylotypic period. *Genome Res.* (2014).
8. Schurmann, C. *et al.* Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PLoS One* **7,** e50938 (2012).
9. Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* **32 Suppl,** 490–5 (2002).
10. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).
11. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS One* **9,** e78644 (2014).
12. Wang, C. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* **32,** 926–932 (2014).
13. Duintjer Tebbens, J. & Schlesinger, P. Improving implementation of linear discriminant analysis for the high dimension/small sample size problem. *Comput. Stat. Data Anal.* **52,** 423–437 (2007).
14. Comte, A., Roux, J. & Robinson-Rechavi, M. Molecular signaling in zebrafish development and the vertebrate phylotypic period. *Evol. Dev.* **12,** 144–156 (2010).
15. Yang, H. *et al.* Deep mRNA Sequencing Analysis to Capture the Transcriptome Landscape of Zebrafish Embryos and Larvae. *PLoS One* **8,** e64058 (2013).
16. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203,** 253–310 (1995).
17. Xue, L. *et al.* Global expression profiling reveals genetic programs underlying the developmental divergence between mouse and human embryogenesis. *BMC Genomics* **14,** 568 (2013).
18. Yanai, I., Peshkin, L., Jorgensen, P. & Kirschner, M. W. Mapping Gene Expression in Two Xenopus Species: Evolutionary Constraints and Developmental Flexibility. *Dev. Cell* **20,** 483–496 (2011).
19. Tan, M. H. *et al.* RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. *Genome Res.* **23,** 201–216 (2013).
20. Wilson, C. L. & Miller, C. J. Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics* **21,** 3683–3685 (2005).
21. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43,** e47–e47 (2015).
22. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264 (2003).

23. Maekawa, M., Yamamoto, T., Kohno, M., Takeichi, M. & Nishida, E. Requirement for ERK MAP kinase in mouse preimplantation development. *Development* **134,** 2751–2759 (2007).
24. Yosef, N. & Regev, A. Impulse control: Temporal dynamics in gene transcription. *Cell* **144,** 886–896 (2011).
25. Bar-Joseph, Z., Gitter, A. & Simon, I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Publ. Gr.* **13,** (2012).
26. Sanchez, A., Choubey, S. & Kondev, J. Regulation of noise in gene expression. *Annu. Rev. Biophys.* **42,** 469–491 (2013).

Figure 1. The Hourglass Model of Development



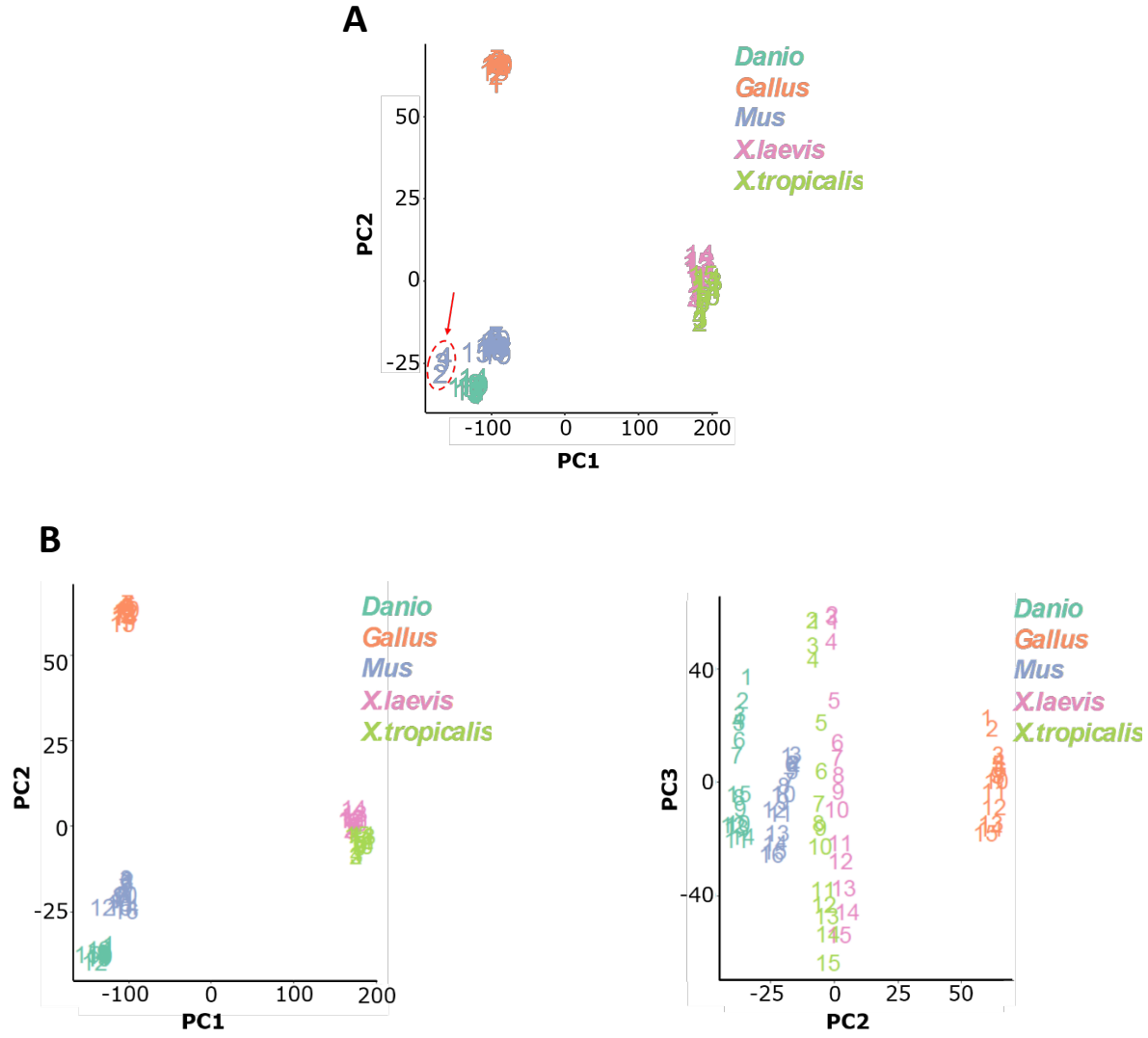
The hourglass model of development (A; Irie and Kuratani (2011)) posits an intermediate "phylotypic" period in vertebrate development during which there is lower interspecific variation in anatomy and gene expression (the narrow middle of the hourglass) than at the beginning or end of development (the bottom and top of the pictured hourglass, respectively). Other possible hypotheses (B) are that there is a steady increase in variation over time (early conservation) or that no significant pattern exists (null hypothesis).

Table 1. Description of data sets collected for meta-analysis.

Publication	Organism	Expression profiling platform	number of replicates per time point	number of time points sampled	range of time points sampled	range of time points (% of development time)
Yang, et al., 2013	*Danio rerio*	Agilent 2100 Bioanalyzer*	1	9	2 hpf to 72 hpf (7 dpf omitted)	2-75%
Roux and Robinson-Rechavi, 2008	*Danio rerio*	Affymetrix GeneChip Zebrafish Genome Array	2	10	6 hpf to 4 dpf (0.25 hpf and 5, 14, 30, and 90 dpf omitted)	6-100%
Irie and Kuratani, 2011	*Gallus gallus*	Affymetrix GeneChip Chicken Genome Array	2	15	HH 1 to HH 38	8-100%
Irie and Kuratani, 2011	*Mus musculus*	Affymetrix GeneChip Mouse Genome 430 2.0	2	8	TS 11 to TS 26	39-95%
Xue, et al., 2013	*Mus musculus*	Affymetrix GeneChip Mouse Genome 430 2.0	3	11	TS 1 to TS 27	5-100%
Yanai, et al., 2011	*Xenopus laevis*	Agilent (custom microarray design)	3	15	Stage 2 to stage 40	2-100%
Yanai, et al., 2011	*Xenopus tropicalis*	Agilent (custom microarray design)	3	13	Stage 2 to stage 40	2-67%
Tan, et al., 2013	*Xenopus tropicalis*	Illumina HiSeq 2000*	Varies (1-3)**	19	Stage 2 to Stages 44-45	2-100%

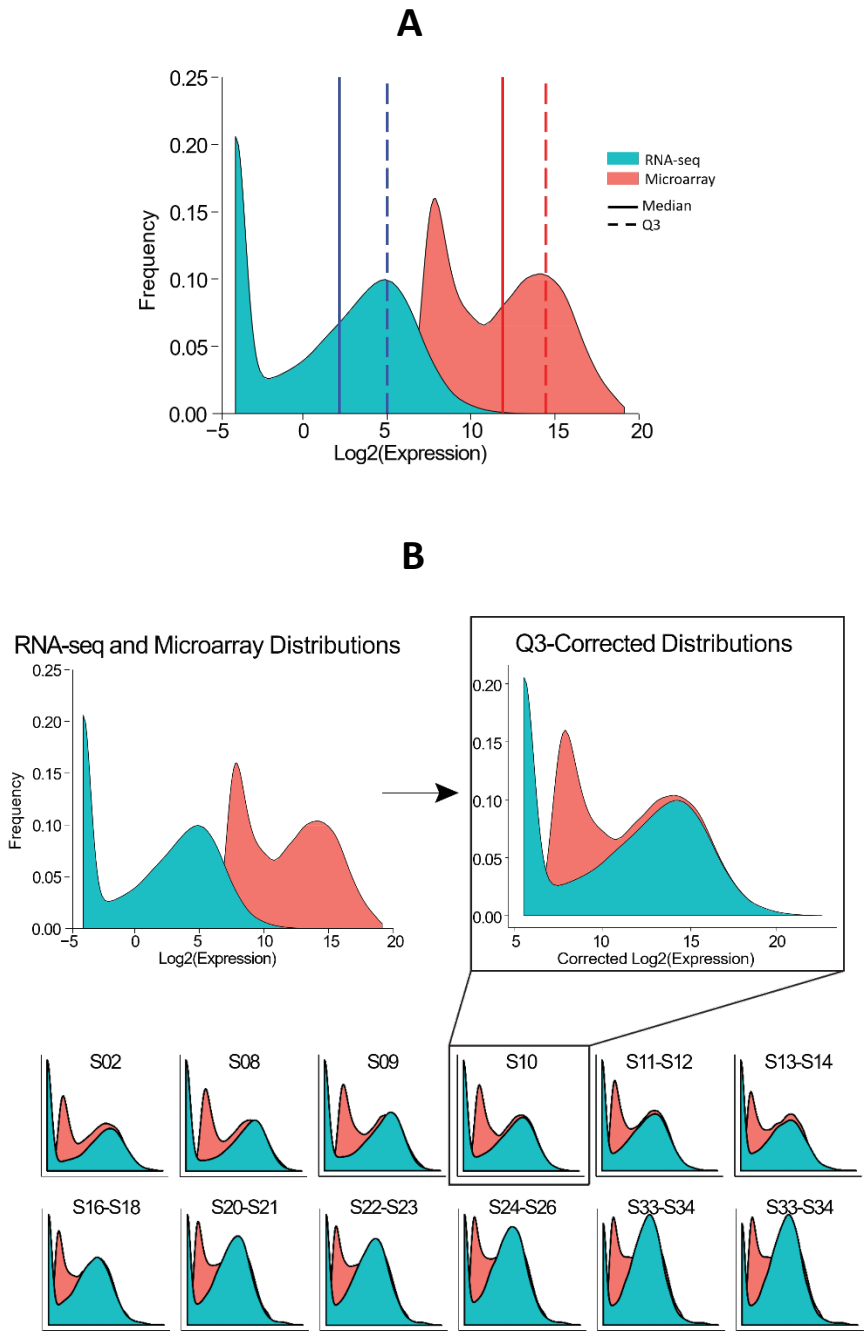
Note: In *X. tropicalis*, each time point includes all data within the same range of developmental stages. For example, X. tropicalis time point #6 includes Yanai, et al. (2011) data from stages 13 and 14 (separate samples) and Tan, et al. (2013) data from stages 13-14 (pooled sample). All other species sample one developmental stage per time point. In *D. rerio*, the 15min time point was elminated due to high amounts of maternal transcript. All data after 4 dpf was excluded based on completion of embryogenesis.
*RNA-seq data set
**See Supplemental Table 1

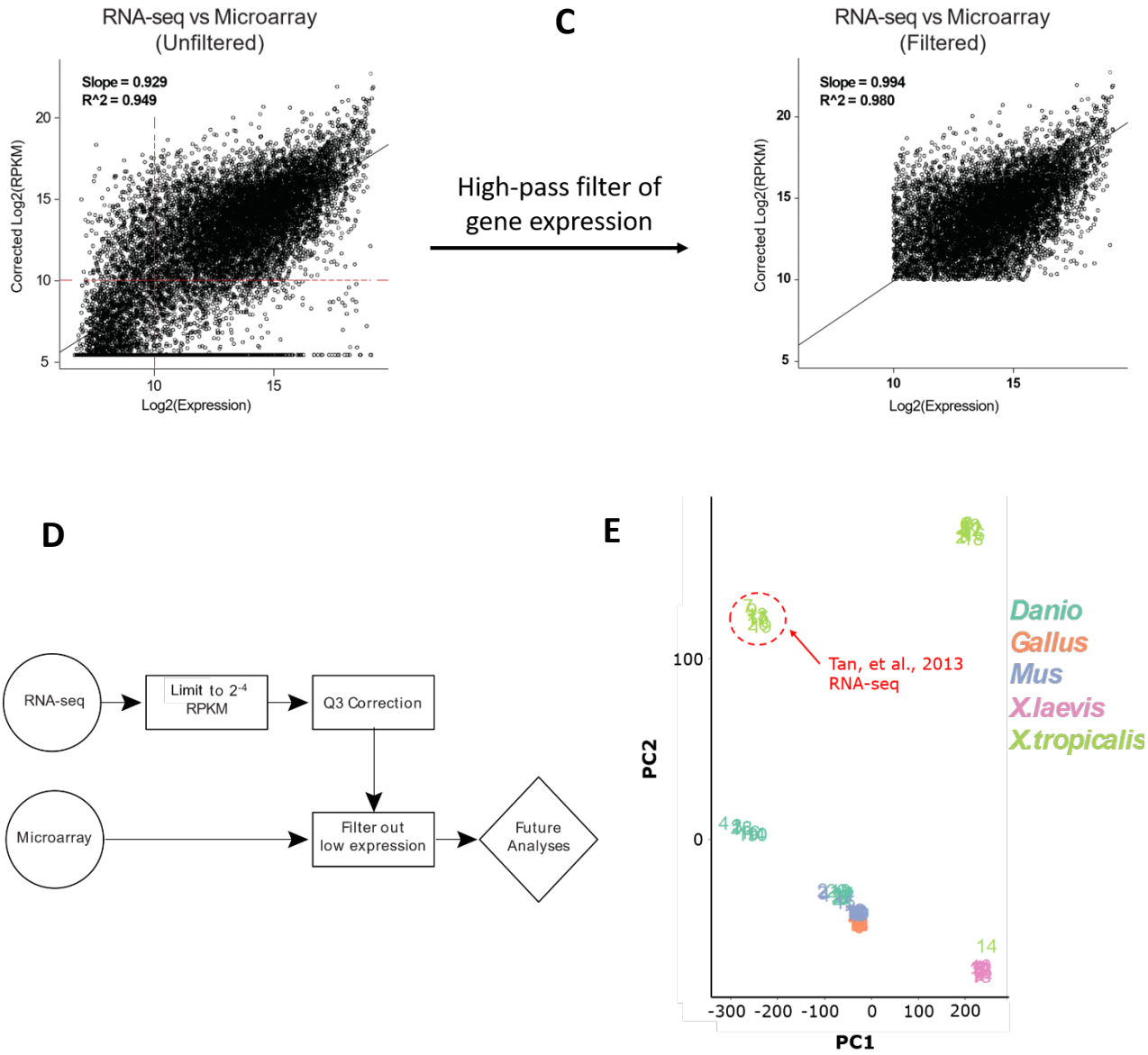
Figure 2. Interspecific PCA exposes batch effects in *Mus musculus* data and identifies species and developmental time as the greatest sources of variation.



(A) Modeling of interspecies variation by PCA revealed that, in the first principal component (PC1), a *M. musculus* data set encompassing TS2, TS3, and TS4 contributed a significant batch effect to overall variation, showing an effect size greater than that between the remaining *M. musculus* time points and the cluster of *D. rerio* time points. (B) After removal of this data set, interspecies variation clearly contributes the most to overall variation in the data and encompasses PC1 (83.02%) and PC2 (4.75%). Progression through development is the next-largest source of variation, encompassing PC3 (3.06%).

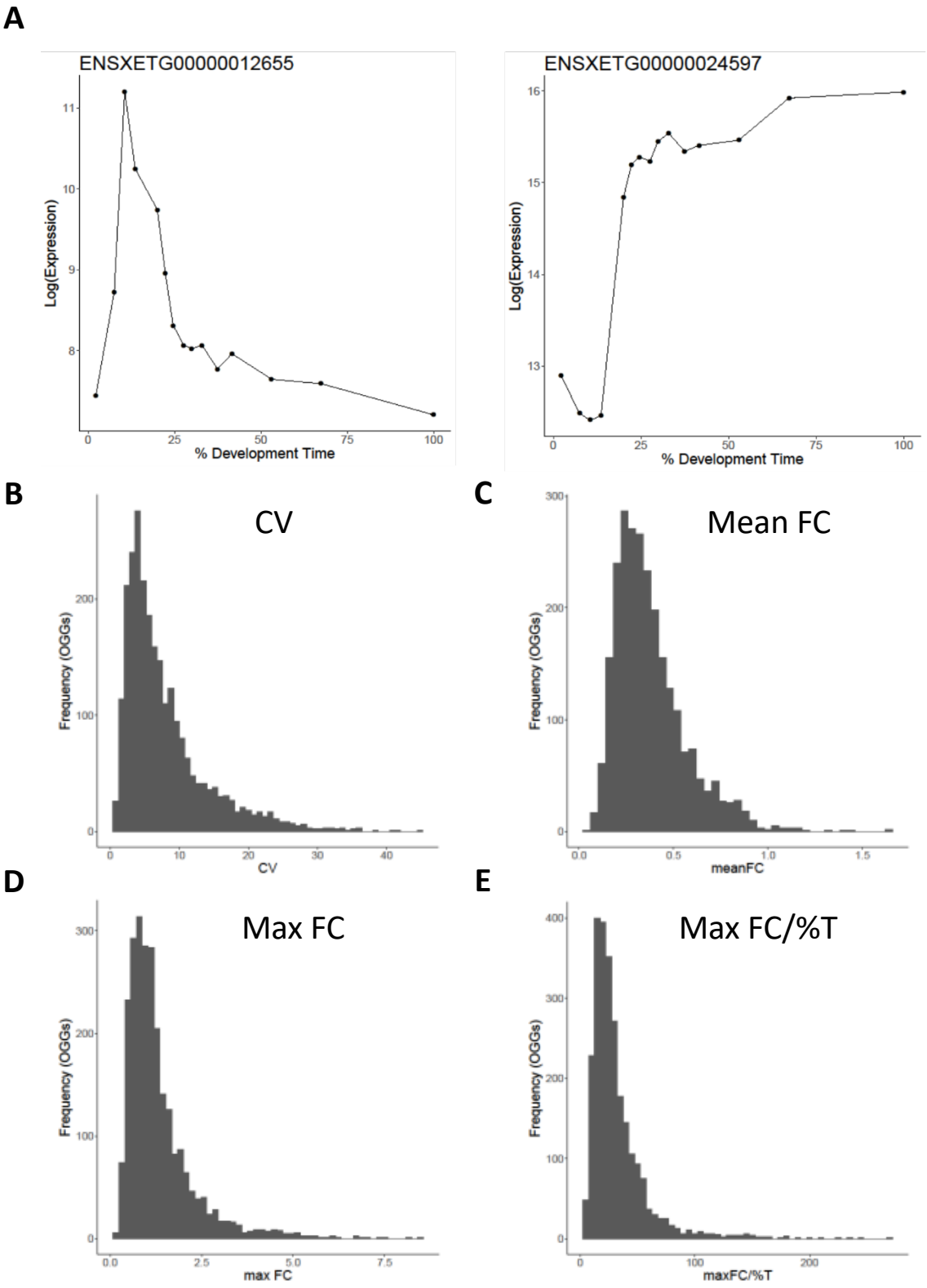
Figure 3. Interspecific PCA exposes batch effects in *Mus musculus* data and identifies species and developmental time as the greatest sources of variation.





(A) After limiting the dynamic range of RS expression to a minimum of $\log_2$(RPKM) = -4, RS and MA showed bimodal distributions of gene expression values within each stage of *X. tropicalis* development (Stage 10 shown). The third quartile (Q3) measures the central tendency of the higher expression mode, so (B) a scalar correction factor was added to align the RS and MA distributions at Q3. Q3 correction resulted in overlap of the higher mode at all 12 shared time-points. (C) Finally, a high-pass filter was applied to data in Stage 10 to select only genes in the higher mode (Log_2(Expr) > 10) in both platforms. The filter improved correlation and slope of regression. (D) A summary of the MA-RS integration pipeline is shown. (E) However, RS data still clustered separately from MA in interspecies PCA, suggesting significant differences despite high correlation.

Figure 4. Naïve descriptive metrics score importance of gene/OGG expression patterns in long developmental time-series.



F. *h* = 0.50

	CV	mean FC	max FC	max FC/%T
CV	1264	989	968	983
mean FC	989	1264	1042	1025
max FC	968	1042	1264	966
max FC/%T	983	1025	966	1264

G. *h* = 0.75

	CV	mean FC	max FC	max FC/%T
CV	632	431	419	401
mean FC	431	632	449	436
max FC	419	449	632	406
max FC/%T	401	436	406	632

H. *h* = 0.90

	CV	mean FC	max FC	max FC/%T
CV	253	158	147	137
mean FC	158	253	155	152
max FC	147	155	253	140
max FC/%T	137	152	140	253

Important gene/OGG expression patterns were considered to be an impulse or sustained response. (A) Temporal expression of *X. tropicalis* genes ENSXETG00000012655 and ENSXETG00000024597 are shown vs. percent of development time to demonstrate impulse response and sustained response, respectively. (B-E) Four expression metrics were conceived to naïvely capture these patterns from temporal expression data (CV, mean FC, max FC, and max FC/%T). All four metrics show a right skew, suggesting higher scoring for a population of highly time-variable genes. (F-H) To assess agreement between metrics, OGGs receiving a score above a threshold quantile *h* in *X. tropicalis*, and the size of each pairwise overlap between metrics was computed for *h* = 0.50, 0.75, and 0.90.

Table 2. Gene expression data sets suffer from small sample size/high dimensionality.

Species	# of replicates per time-point	# of genes after pre-processing
Danio rerio	2	8729
Gallus gallus	2	7934
Mus musculus	3**	17403
Xenopus laevis	3	16260
Xenopus tropicalis	3	16260

Supplemental Table 1. Sample sizes of data published by Tan, et al. (2013).

stage number(s) covered in time point	number of replicates
2	1
8	1
9	2
10	2
11, 12, 11-12	3
13-14	1
15	1
16, 16-18	2
19	2
20-21	2
22-23	2
24-26	2
28	2
31-32	2
33-34	2
38-39	1
40	2
41-42	2
44-45	1

Supplemental Table 2. Real time at developmental time-points for all species microarray data.

Publication	Species	Stage of development	Time of sampling	% development time (%T)
Roux and Robinson-Rechavi, 2008	*Danio rerio*	shield	6 hpf	6%
		75% epiboly	8 hpf	8%
		90% epiboly	9 hpf	9%
		bud	10 hpf	10%
		5-somite	12 hpf	13%
		14-somite	16 hpf	17%
		prim-5	24 hpf	25%
		32 hpf	32 hpf	33%
		long-pec	2 dpf	50%
		4 dpf	4 dpf	100%
Irie and Kuratani, 2011	*Gallus gallus*	HH1	24 hpf	8%
		HH2	30 hpf	10%
		HH4	42 hpf	13%
		HH6	48 hpf	15%
		HH8	51.5 hpf	17%
		HH9	55 hpf	18%
		HH11	66.5 hpf	21%
		HH14	75.5 hpf	24%
		HH16	78 hpf	25%
		HH19	102 hpf	33%
		HH24	132 hpf	42%
		HH27	150 hpf	48%
		HH32	204 hpf	65%
		HH34	216 hpf	69%
		HH38	312 hpf	100%
Irie and Kuratani, 2011	*Mus musculus*	TS01	1 dpf	5%
Xue, et al., 2013		TS09	6.5 dpf	34%
		TS11	7.5 dpf	39%
		TS13	8.5 dpf	45%
		TS15	9.5 dpf	50%
		TS16	10 dpf	53%
		TS17	10.5 dpf	55%
		TS19	11.5 dpf	61%
		TS20	12 dpf	63%
		TS21	13 dpf	68%
		TS22	14 dpf	74%
		TS23	15 dpf	79%
		TS24	16 dpf	84%
		TS25	17 dpf	89%
		TS26	18 dpf	95%

		TS27	19 dpf	100%
Yanai, et al., 2011	*Xenopus sp.*	S02	1.5 hpf	2%
		S08	5 hpf	8%
		S09	7 hpf	11%
		S10	9 hpf	14%
		S12	13.25 hpf	20%
		S13	14.75 hpf	22%
		S14	16.25 hpf	25%
		S16	18.25 hpf	28%
		S18	19.75 hpf	30%
		S20	21.75 hpf	33%
		S23	24.75 hpf	38%
		S25	27.5 hpf	42%
		S30	35 hpf	53%
		S33	44.5 hpf	67%
		S40	66 hpf	100%

hpf = hours post-fertilization

dpf = days post-fertilization

Biography

Pranav S. Bhamidipati was born in Hyderabad, India on 19 February 1995 and moved around the northeastern and midwestern United States with his family from 1996 before moving to Houston, Texas in 2001. He enrolled in the Plan II Honors program at the University of Texas at Austin in 2013, pursuing degrees in Plan II Honors and Biochemistry. In college, he pursued summer research projects at Baylor College of Medicine in Houston, Texas, and won multiple awards for his work at UT Austin in the lab of Dr. Hans Hofmann on the computational study of evolution and development. The summer after his junior year, he studied the mechanism of mycobacterial persistence to fluoroquinolone antibiotics at the Institut Pasteur in Paris, France, under Dr. Giulia Manina. Mr. Bhamidipati graduated Phi Kappa Phi in 2017 as a Plan II Distinguished Graduate. He plans to join the MD/PhD dual degree program at the USC Keck School of Medicine and California Institute of Technology this fall.